Trust and Responsibility

Digital Systems from a Capacity-Based Perspective

Andreas Kaminski, Marcus Düwell, Philipp Richter

Abstract *Decisions in companies, public administrations, and political institutions are increasingly influenced by complex computational models. In response, numerous ethical guidelines and standardization approaches – often referred to as "AI ethics" – have been developed. Many of these approaches focus primarily on the properties of the models (such as fairness, reliability, transparency, or privacy). The prevailing assumption is that only if these properties are met can the use of such systems be considered responsible and trust in them justified. However, what is often overlooked is that this approach implies an important precondition: individuals and organizations must have the capacity to assess these systems in terms of their fairness or reliability. Given the increasing complexity – and thus opacity – of many models, it is questionable whether this precondition is actually fulfilled. This question is therefore of fundamental importance for legal and moral discourses on AI. In our article, we introduce the idea of a capacity-oriented approach. We show how virtue-ethical considerations and moral principles aimed at agency and autonomy can provide the foundation for this approach. Building on these reflections, we propose a revised requirement for digital governance and raise the question of how governance can be designed to make systems responsible in such a way that their use can be considered trustworthy.*

1. The Usual Approach and its Critical Prerequisite

In many discussions surrounding digitalization and artificial intelligence, we encounter demands for *trustworthy* systems that have been designed *responsibly* (Pekka et al. 2018; Floridi 2019; Thiebes et al. 2021; AI Act 2024). Attention here primarily focuses on the normatively relevant properties of such systems, such as being fair, reliable, or health-promoting. According to this view, a system is considered trustworthy if it possesses certain properties, for instance, being just and reliable (Bisconti et al. 2024; AIEI Group 2020). If it exhibits these properties, it can also claim to have been responsibly designed, and its deployment may potentially be regarded as justifiable. However, this line of thought explicitly or implicitly presupposes that

individuals, communities, or organizations are able to relate the functionalities and performances of such systems to normative considerations relevant to them (e.g., values, beliefs, principles, or rights).

> **Thesis 1:** Reasonable trust in systems and the acceptance of responsibility for these systems require the capacity to evaluate whether such systems are worthy of trust and can be used responsibly.

A person, community, or organization that wants to evaluate whether a system is just or reliable, promotes health or autonomy, must understand (1) *how that system works*, at least in principle, and must also understand (2) *how the system relates to normative aspects*. The twofold prerequisite is essential for the evaluation of technical systems in general, and digital systems in particular.

This raises two questions. The first is a *factual* question: Is this prerequisite fulfilled with regard to a specific technical system? And if so, for which actors and to which degree? This question can, for example, be examined within the context of sociological studies. The second is a *normative* question: What conditions should be in place to enable individuals or organizations to evaluate technical systems? How can we ensure that this critical condition is fulfilled? How should systems be designed to support individuals or organizations to achieve this capacity? And furthermore, how can organizational environments be created that foster the development of such capacities and prevent technical developments that hinder them? Both questions lead directly to the issue of digital governance on the one hand and an enabling-capacities approach on the other hand.

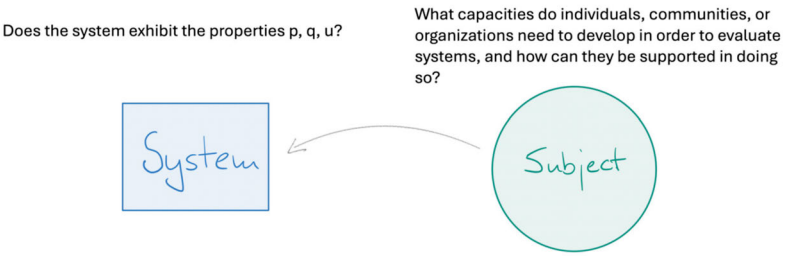
Consider a system for granting loans or a medical system that diagnoses illnesses and recommends therapies. We assume here that both systems operate based on machine learning methods. For individuals or organizations to be able to evaluate loan decisions or medical diagnoses, they must first understand which factors were decisive for the loan approval or the diagnosis. However, this alone is insufficient; they must also be able to assess the quality and appropriateness of the decisions. Imagine that a person's loan application has been rejected, or that radiation therapy is recommended due to a tumor diagnosis. To evaluate such decisions or recommendations, individuals must understand whether they are well-founded, fair, or beneficial for health. This brings us to the role of the capacities for understanding and evaluating. Beyond this, the question arises whether the organizations operating these systems enable reasonable interactions with them. Can objections, for example, be raised and addressed appropriately if a loan is rejected without appropriate reasons? Do the involved medical experts have sufficient time and knowledge to engage with the systems and its decision, to question it, learn from it, and, if necessary, to reject its recommendation? Such questions pertain directly to the digital governance of organizations.

2. The Enabling-Capacities Approach

Here, this book proposes a shift in perspective: Until now, ethical and legal debates – especially in the context of AI – have primarily focused on directly evaluating the properties of technical systems; judging whether these systems meet certain preferred conditions. However, we argue that to fully grasp what it means to regard certain system properties as good, better or worse, we must first take a step back. Before we may focus on particular properties, we need to reflect on the task of evaluation itself.

Our claim is this: Whenever a system is evaluated, there is always someone who is making that evaluation through a value judgment. Evaluative judgments, when made on a reasonable basis, presuppose that certain conditions are met. The subject who makes the judgment must possess specific capacities and be able to exercise them appropriately. This raises the question of what those capacities are, how they can be cultivated, and what conditions are particularly conducive to their development and exercise. Evaluating technical systems therefore requires *capacities* that are a prerequisite for evaluation; and it raises the question of the conditions that promote, facilitate, impede, or prevent the development and exercise of these capacities.

Figure 1: From a focus on properties to a capacity-based approach.



In principle, the capacity-based approach has always been assumed within the context of examining the properties of a technical system. What we are doing here, first, is to make this seemingly unproblematic premise explicit, since it is by no means always uncritically fulfilled. Second, we are no longer directing our normative questions solely at these properties, but primarily at the conditions under which capacities are developed to identify and evaluate those properties. These conditions, after all, can themselves be shaped, politically demanded, and technically supported. Third, this may lead to considerations that urge to come

up with a more integrated normative view on the normative assessment of those technical systems.

> **Thesis 2:** Not only the properties of technical systems can be subject to ethical, political, and legal evaluation, but also the conditions under which such evaluations can take place. What is evaluated, then, are the conditions under which capacities are developed to identify and assess these properties. After all, these conditions can themselves be shaped, politically demanded, and technically supported.

Although, as mentioned before, this shift in perspective is not without precedent. One of our most important sources of inspiration lies in philosophical ethics of prudence dating back to Aristotle in ancient times. Within this tradition, as we can say, attention is paid to the capacities required for making appropriate value judgments. At the same time, consideration is given to the conditions conducive to prudent and ethical action.

Some words about *ethics* may be necessary here. Ethics is here not understood as a set of customs or rules of behavior but as a philosophical discipline – which is a common view since ancient times. As a philosophical enterprise, ethics forms a discipline that in a systematic way reflects on questions like 'What shall I or shall we do?' or 'How shall I lead my life?'. In that sense, morality, moral norms, duties, rules, or principles are the topics on which ethics reflects. Philosophical ethics systematically aims to understand what is so specific about moral norms and values, how we can speak about them, and whether those moral requirements are only expressions of subjective views – only opinions – or whether moral requirements of right actions can be justified in intersubjective context. But moral norms and principles are not isolated in human praxis. We are not only confronted with moral questions but in normal life we ask as well, how to live one's lives? What is important for us in life? What makes us happy? These are questions where some variety of answers are possible. While some would be ultimately happy to play several hours of football a day, for others that would be a nightmare. Nevertheless, there seem to be some considerations that seem to be quite generally valid. Being never able to fulfill their dreams and aspirations is something that makes people unhappy, independent of what the dreams and aspirations may be. Permanently ignoring one's talents and not developing them seems not a good idea for human beings in general – even so it may not be immoral to do so. Thus: There is a dimension of human life that does not coincide with morality but about which ethics has something to say.

This field is often called an 'ethics of prudence'. Today, prudence or prudential reasons are often seen in direct opposition to morality, since prudence is understood to deal with what one could ideally achieve for oneself. This is because prudential reasons seem to be derived from mere instrumental means-ends-calculations. If we say a person behaves prudently, we seem to articulate the suspicion that the

person aims at achieving whatever selfish ends they might have – and the smart guy is not necessarily the good guy. On this view, ethical theory must establish justifiable constraints to counter the prudent pursuit of self-interest (Kaspar 2011: 311 et seq.). However, this is only half of the story, prudence was and is seen also as a complex intellectual and practical virtue, being bound to certain moral values or complementing them practically (Annas 1995: 245, 251; Luckner 2005). There is a philosophical tradition of diverse thinkers such as Aristotle, Machiavelli, or Descartes that can be called the "ethics of prudence" (Luckner 2011; Benner 2010; Cimakasky and Polansky 2012). This tradition is especially helpful for tackling the challenging blend of descriptive and normative questions in applied ethics (Richter 2018: 47–51). This ethical approach does not necessarily conflict with principle-based ethics or universal moral norms, rather it complements them by considering the processes and criteria through which individual subjects deliberate, judge, reason, and act according to ethical reasoning (Hubig 2007: 140; Fremstedal 2018). We will come back to this relationship between ethics of prudence and morality below.

2.1. Applied Ethics as an Enabler of Ethical Reasoning (Hubig)

German philosopher Christoph Hubig developed an approach to applied ethics which heavily draws from ethics of prudence (Hubig 2007: ch. 5; Hubig and Richter 2015). By descriptively acknowledging moral pluralism, both, in practical reality and in moral theories, Hubig draws our attention to the common ground of all ethical theory: All ethical approaches are focusing on different aspects of free human action. Here, freedom is not understood as human beings acting without any constraints, like external restrictions and forces which they do not have chosen, or internal limitations, like inhibitions, fears, or lack of motivation. However, we must assume that human beings can reflectively relate to their behavior and ask questions like 'What can I do? What do I want to do? What am I obliged to do?' Freedom in that sense is necessarily presupposed since ethics deals with qualifying our action as prudent, right, or good. All ethical approaches must presume that there is free action since otherwise all ethical reasoning would be futile and practically irrelevant (Hubig 1995: 113). This holds for moral theories about universal duties and for prudential approaches.

Ethics of prudence starts, as we can say, with one general assumption about practical freedom: It is not within our power to do everything as we please, but neither is everything beyond our control. Practical freedom is understood as a reflective behavior of setting oneself in relation to identified determinants, thereby gaining clarity on the available options for action (this also could be called 'to orientate oneself' in the sense of getting oneself ready to act, see Luckner 2005: 9–30). Here, freedom is both an ability and its procedural actualization in deliberation, which involves learning about what one's abilities and options are, and thereby getting one-

self ready to choose what is normatively good in the long run. It is an epistemic subtask of prudent judging and acting to take precautionary measures against real or potentially upcoming dependencies (a lack of practical freedom) and to extrapolate these from empirical knowledge as well as to actively obtain the necessary knowledge about immediate and future circumstances of action.

However, prudential ethics does not simply assume freedom as a matter of fact. Freedom is vulnerable, can be restricted, and is endangered. Therefore, it asks how freedom can be preserved or enhanced in practice. A central aim of all prudent thinking and a general good to agents is therefore the long-term preservation and increase of the ability to act because, for instance, according to Aristotle, human activity in the sense of the associated virtues must be considered intrinsically valuable from a practical perspective. Without going into detail about a foundational argument, it is quite plausible to take the 'ability to act' as a general good of all prudent deliberations: in general, it is true that situations need to be avoided in which one is forced to react without free deliberation, having only little or no other choices, since it is bad to be unable to do what one really wants or what would seem reasonable. We can say true prudent deliberations are only those that reach their specific ends and at the same time allow for further prudent deliberations in the future. Prudence could be seen as a capacity-saver and enabler. As far as possible, a prudent person would avoid severe restrictions or the loss of the ability to act. Known or deducible examples of the severe limitation or loss of the ability to act are for instance: severe psychological or physical dependencies (such as addictions), but also lack of income, as well as socially or naturally induced disasters (such as wars or flooding), technical constraints (*Sachzwänge*) as well as lack or failing of infrastructure. All of those should be prudently avoided by taking precautionary measures even if this is often only realizable via collective efforts.

> *Let us imagine, for example, that it is a legacy system that has been maintained by different people over long periods of time and is poorly documented. Major changes to such a system may either appear to be unfeasible, with the result that the only option is to try to keep operations running. Or, to give another example: a system has become so dominant that there are no real alternatives on the market. Cases such as these are to be avoided by acting prudently, as they severely limit or even exclude the possibility of prudent behavior in the future.*

Now, to avoid all these anti-prudent restrictions could be seen as normatively demanded by rational egoism being in favor of an individual's own agency to the possible detriment of others. However, this is not necessarily the case. We can consider the ability to act as a good while we can see prudence as a way of reasoning to maintain this good. If we see it like that, prudence generates the preconditions for possible moral action in the long term, since without the ability to act there is no ethical reasoning and no moral action according to it. So, to achieve moral action according

to universalist morality there must be prudent thinking (Hubig 2007: 129, 131 et seq.; Kaspar 2011: 320 et seq.). We find that line of thought, for instance, also in Kant's universalist moral theory when he speaks of an 'indirect' duty to "secure one's own happiness" (AA IV: 399), since a lack of practical possibilities due to crisis, poverty, fear, or dependency and so on "might become a great temptation to transgression of duty" (AA IV: 399). So, moral reasoning, even according to principle-based ethics like Kant's, involves provision and precautionary measures to sustain the ability of moral action. Here, we may suspend the question of whether true prudence is already bound to moral virtue, as for instance Aristotle took it (in contrast to Hobbes and Kant). For us, it is sufficient to assume that prudent thinking is bound to the end of sustaining agency in persons, groups, or states, which is a practical precondition to deliberate about actions also in an ethical way. The point is, first, there must be at least some options to form decisions and actions according to them, then, in a second logical step, these can be evaluated from an ethical point of view. Without options, no actions. And without actions, no object of ethical reasoning. In this respect, actions (or system properties) may be judged as morally good or bad. However, to form these judgements we also have to admit that it is good that the world contains at least some actions (or persons judging system properties); otherwise, ethical reasoning would be futile and meaningless. For sure, prudent action is not the whole of ethics, since it does not offer founding arguments for principles holding universal moral norms. However, universal moral norms or goods can only be fulfilled if their addresses are able to act now and in the future.

To come back to Christoph Hubig's approach to applied ethics, the common ground of all the diverse ethical reasoning according to a standpoint of ethics of prudence is to sustain the ability to deliberate and to act freely. If this is combined with foundational theories of universal morality, e.g., Kantian ethics of autonomy, we find two complementing aspects of free agency in agents being able to ethically reflect under conditions of system induced chances, dependencies, and restrictions (Hubig 2007; Hubig and Richter 2015).

2.2. Option and Legacy Values

Since freedom exists in relating oneself to not or only hardly changeable determinates, and in choosing in the remaining leeway of options considering some of them as better or worse (cf. Raz 1988: ch. 14), all agents need a variety of options to choose from, otherwise there exists no practical freedom. We can call this a view of practical freedom "from outside". However, considered "from inside", agents also need to be able to consider themselves a constant self over time and during different sequences of action, for if my standards of evaluation changed from one moment to the next, no action extending beyond the immediate moment would be possible. This capacity – being the author of one's actions – depends on social preconditions

(ibid.). For instance, if a society allows violence or disrespect, it is nearly impossible to develop long-term plans or to inculcate agency supporting virtues like practical self-efficiency, self-respect, or trust in one's own abilities.

The preservation of the ability to act, regarded as the highest aim of prudential ethics in this sense, according to Hubig, presupposes two fundamental types of values:

1) "option values": the preservation of options for action;
2) "legacy values": the preservation of being a subject, namely the ability of cultivating and maintaining a personal identity that informs a way of life from past to present (Hubig 2007: 141–145).

Shaping technology requires options. Without at least two options (a or b), there is no space for the design of technology. When no options are given, for instance, because practical constraints exclude alternatives, then evaluations are not feasible or meaningless. One can still arrive at the conclusion that a certain technology is 'bad', but this judgment may remain inconsequential.

> *Let us imagine, for example, a software that has become the standard, and which one considers poor in numerous respects (it is too expensive, does not adequately protect data, and restricts possibilities for action). The company that produces the software is aware of this criticism; however, it does not respond, or responds only selectively to the deficiencies, because it knows its status as an (industry) standard. The same applies to models for evaluating creditworthiness that have become quasi-standards, as well as to development paths (such as in electricity or mobility) that can no longer be easily changed.*

Furthermore, the assessment of technologies requires that subjects can relate to technologies. This means that subjects must know that and how they are affected by these systems; for this, they must be aware of the effects of the system. However, this is not the case if the effects remain below the threshold of perception or cannot be understood. Furthermore, subjects must not be influenced by the systems in such a way that their standards become incoherent.

It is important to see that in this perspective the mere existence (and maintenance) of options to choose from and the preconditions for generating subjects (e.g., family structure, social services, education at schools or universities) is prudently good and indirectly morally required even if these possibilities will not actually be used with specific benefit in the near future. The mandatory provision to keep these possibilities in stock does not necessarily involve concrete action plans of how to make use of them; this should be left to the individuals of whom ethical reflection in one or the other way is expected.

Let us imagine a system that assesses the creditworthiness of individuals without their knowledge that such a system is being used for this purpose. They also may be unaware of what data is taken into account and how it factors into credit decisions. Such a system, for instance, could monitor the browsing history of individuals, the history of their addresses, and much more without them being aware of this surveillance or without them understanding its impact on the credit decisions being made. This is why legacy values are required in order to evaluate technological systems.

Option and legacy values are meta-values (not specific rules or principles) functioning as what in ancient Rhetorics is called "topoi". Topoi are criteria which need to be considered always when dealing with certain topics or problems. However, they always need specification based on empirical knowledge about the real-life situation now dealt with (Hubig 1990: 134, 140 et seq.; Richter 2017: 190, 195 et seq.). If we follow Hubig's approach, we may accept the pluralism of moral approaches and diverse ways of ethical reasoning (which is one of the preconditions of free agency in a modern society) but follow a perspective for applied ethics which is valid universally: A technical system, may it be in health care, internet-based technologies, or traffic infrastructure, could only be morally acceptable or good if users are and kept able to relate themselves to the systems infrastructure, to gain practical freedom, and the possibility for ethical reflection. According to option values, technical systems must grant the users a scope of possibilities for creative redesign, rededication, critique, discussion, or rejection of intended ways of use. If there is little opportunity for choice in the sense mentioned before, e.g., if the determinants or system functions cannot be identified and no alternate ways of use are visible, then there is little chance for users to reflect on their behavior and that of the system in a broad and ethical way.

However, how to achieve this in a technical system is not straight forward like an easy deduction, since there are conflicting goals, since technical systems relieve and facilitate procedures which otherwise needed to be done manually by human action. Therefore, a prudently wise balance between giving up freedom by making use of system infrastructure and maintaining freedom while relying on a system is required (Hubig 2007: ch. 6; Luckner 2005: 164 et seq.). This also holds for further specifications of the prudential approach. Topoi of action enabling capacities need to be developed, both in a theoretically and empirically informed way. In a next step these considerations need to be specified in detailed studies about the technical systems in question. If we consider algorithmic based digital systems, then the topoi understandability and controllability need to be considered since these capacities function as preconditions of users' practical freedom under the restrictions of these types of systems. For without the ability to understand technical systems, it is impossible to recognize how they shape and transform practices and decision-mak-

ing contexts; without the controllability of technical systems, the systems cannot be shaped or steered.

2.3. Prudence and Morality (Aristotle and Kant)

We now have presented some reasons why the freedom of agents to relate to technologies in a free and controlled way can be justified from a broadly speaking Aristotelian approach. One may wonder whether these considerations were only plausible if one subscribes to an Aristotelian concept and whether this is not a type of ethics that is at odds with a modern concept of morality which is more focused on universal moral principles. This is particularly relevant since central modern concepts like human rights and human dignity which form the cornerstone of a liberal worldview are based on those universal aspirations. Philosophically, the most obvious comparison to discuss these questions is the comparison between Aristotle and Kant. In contrast to prudential reasoning, a principle-based approach is often understood to be focused solely on the justification of universal moral principles, thereby neglecting the contingencies of practical reality in which these have to be specified and employed wisely. This applies especially to how Kant is commonly understood. Thus, the difference between context-sensitive prudence and strict universal moral principles seems to be the difference between Aristotle and Kant.

We propose a different take on the relation between Aristotle's idea of prudential reasoning and Kant's moral philosophy. Already in the *Groundwork*, Kant emphasizes that agents have to see themselves under three imperatives, we could call them (in modern terminology): instrumental, eudaimonistic and moral imperatives. Kant-scholarship focused primarily on the moral imperative, which is the only one categorically valid, that is the only imperative that is binding for agents under all circumstances. But Kant stresses that there are three imperatives (see for this interpretation of Kant: Steigleder 2002: 23–58). Whenever an agent is committed to realize goals, he must also be committed to the means which are necessary to reach these goals. If it is the goal of an agent to become a successful guitar player and if she is really committed to this goal, she has to practice regularly, otherwise she cannot claim to be committed to this goal. This is imperative for the agent in a strict sense, an agent that would assume that he is committed to the goal but is not as well committed to the necessary means would not understand himself consistently. This imperative is only in that sense not categorically valid, since it is not necessary for the agent to hold this commitment to this goal, this commitment is contingent; the agent can just decide that she does not want to become a guitar player any longer. But as long as the agent is committed to the goal, he is necessarily committed to the necessary means. Since there can be more than one means appropriate to reach the goal, the agent must also be committed to be able to reflect on the appropriate means and the relevant capacities to realize those reflections. The agent must furthermore

be committed to strive for happiness. If the agent were not under the eudaimonistic imperative, he would not have any reasons to prefer one goal to the other. If the agent did not have any reasons for choosing a certain goal, it would not be plausible why he could be harmed if anyone hindered him to realize his goal since the choice of this goal would be totally random. The ability to form maxims must be guided by this commitment to eudaimonism, otherwise it is not evident how maxims ever could be generated. Thus: the agent is under the eudaimonistic imperative, but this commitment as well is not unconditionally binding but is restrained by the moral imperative that alone is unconditionally valid. There may be conditions where the agent is confronted with duties that are really morally important even so, it will come with restrictions on the happiness of the agent – sometimes the moral demands can really come with far-reaching negative consequences.

Even if in that sense morality can be opposed to the happiness of the agent, the categorical imperative must be embedded in the two other imperatives, otherwise various aspects would not be understandable, here only two should be mentioned: First, Kant stresses that the moral imperative obliges the agent to investigate whether maxims are acceptable for all other agents as well. This already assumes the other has an order of means and ends that is meaningful for the other. Second, Kant stresses that we have a duty to promote the happiness of others. This also assumes that happiness is meaningful for others as well. For Kant, controlling the circumstances of my action and to be able to form an informed judgment is of central importance. The moral imperative is understood as an articulation of our autonomy as rational beings. We should see us as morally bound because the categorical imperative is justified by our consistent self-understanding. This emphasis on the 'capacity to judge' is so to speak the cornerstone of his entire philosophy (Longuenesse 2001). For our context, this is important since the Kantian approach does not only entail that the agent is only capable of deducing concrete norms form general principles. It rather also entails a broad range of capacities that enable the agent to orient oneself in the world. Being able to place oneself in the position of everybody else – as the application of the categorical imperative entails – presuppose that the agent has the imagination to transcend the own position. Being able to form judgments about happiness and morality entails the capacity to imagine what course of actions are possible. In the *Critique of the Power of Judgment* Kant emphasizes the importance of aesthetic judgments for orienting oneself in the world. One could add: Only if the agent can imagine new ways of how she can act, the choice between different courses of action is meaningful (Düwell 1999). In that sense, the focus on a concept of morality as categorically binding norms and principles is dependent on the preservation and cultivation of the power of judgment.

2.4. From Immediate Situations to Long-Term Effects

These considerations are particularly significant in the domain of technology. The necessity of integration (of prudential reasoning and principle-based morality) becomes even more obvious if we think about more complex forms of moral, political, and legal considerations, where we are not only faced with questions about moral duties in single situations, but also where structural questions, long-term effects, and uncertainty about the consequences of actions have to be considered. In all those contexts, the agent first has to try to understand what the current possibilities of actions are, which consequences they may have and how they will affect other agents. All of that presupposes that the capacity to judge is developed and that the agent is in the position to exercise this capacity.

In the line of these considerations, one could emphasize that theories of (human, individual, subjective) rights should also be seen as embedded in such a broader concept of enabling capacities. It has already been mentioned that those rights are the cornerstone of modern societies. The link between rights and enabling-capacities becomes plausible if we consider that rights do not come in isolation. Only in exceptional situations, the protection of only one right is at stake (questions of life and death, torture, emergency situations, etc.). It is, however, much more common that we are faced with competing rights claims or questions where we have to decide about long-term, cumulative, or indirect effects of actions on rights. Quite often there are tensions between the prima facie rights of people to exercise some liberty and rights of others that can only be realized if some liberties are restricted. In all those conflicts questions about hierarchy, urgency and priorities between rights claims are on the table. This leads to questions of consistency in the interpretation within the systematic connection between rights (see Gewirth 1978; Düwell, Graf Keyserlingk and Richter 2025). At least in three respects questions of enabling capacities are relevant. First, commitment to rights implies a commitment to the necessary conditions for the realization of such a right. If we (individuals, groups, states) are committed to a general right to freedom of movement, we must as well be committed to protect and support the conditions that are required to exercise this right. Second, this now implies not only conditions that are necessary for an individual to realize his individual rights but also the collective conditions under which it is in general possible to exercise those rights. In case of freedom of movement this implies at least some form of infrastructure that is required to exercise this right. Third, regarding various rights, it is neither *eo ipso* evident what the necessary conditions for a successful exercise of these rights are nor how the importance of this right can and should be weighed against another right. For that reason, it would at least be required that there are procedures and decision rules which are required to form an informed and collectively accepted decision. All the three dimension show

that there are enabling capacities required under the assumption that one has some commitment to certain rights.

3. The Role of 'Understanding' and 'Explaining'

The evaluation of systems, whether in specific situations or from a long-term perspective, is challenging due to their partial opacity. This opacity first pertains to their internal functioning, which we can refer to as *model opacity* (Humphreys 2009; Beisbart 2021; Burrell 2016).[1] However, there is a second type of opacity that has not yet been explicitly addressed or conceptualized: Opacity can also extend further to include the consequences and side-effects of these systems, as well as the quality of those outcomes. This can be termed *pragmatic opacity*. This opacity does not concern the model itself, but rather the effects of the model in (real-world) application contexts.

> *A simple example would be a web service (e.g., web search) that adapts the selection of information to the users (modeling user preferences). Model opacity would occur if, for instance, we cannot anticipate whether, how, and why a (slight) change in input data would alter the model's results. Suppose a user is 35 years old, female, and lives in Berlin. The model is opaque for us if we cannot comprehend (a) whether, (b) how, and (c) why something would change if she were 33 years old or lived in Munich instead.*
>
> *Pragmatic opacity, in contrast, arises if the user does not recognize (a) that and how the system adapts the information selection for her, or (b) how appropriate (or beneficial) this selection is. Clearly, it is essential not only to identify these different explanatory factors but also to weigh them, particularly regarding their significance for specific (material, fundamental, ethical) values.*

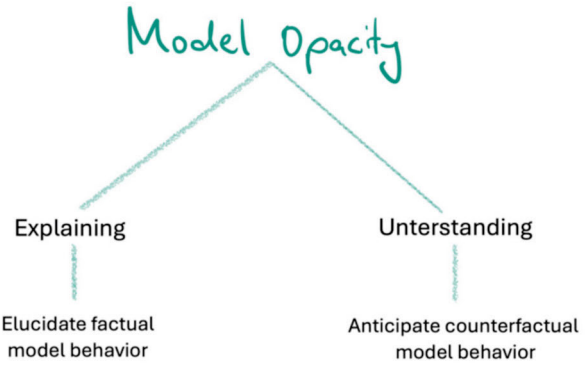
In the case of model opacity, even if the mathematical function of each element (neurons, cost functions, hyperparameters, etc.) can be examined and traced, this alone does not yield an understanding of the overall behavior of the model – as evidenced by the fact that typically one cannot anticipate the behavior of the model, even with only minor adjustments (Kaminski et al. 2018). Pragmatic opacity, on the other hand, refers to situations where the effects of a system on its environment cannot be easily understood or reliably evaluated. This concerns not only the direct effects produced by the system but also the long-term social, political, and other implications associated with its use.

1 This is usually referred to as "epistemic opacity". However, this term is typically understood to refer only to the opacity of models, and not to their pragmatic opacity. For this reason, we prefer the term "model opacity" over "epistemic opacity."

Current research primarily focuses on minimizing model opacity, while neglecting the realm of pragmatic opacity. As a result, the significance of models for their respective fields of application is not adequately considered. (Pragmatic opacity means precisely that one cannot assess the quality of system outcomes; quality with respect to the fields of action deemed relevant.) The concepts of 'explaining' and 'understanding' thus appear from two distinct perspectives: firstly, in relation to the model and the previously discussed model opacity; and secondly, regarding the significance of systems in terms of how they transform our practices; that is, with respect to pragmatic opacity.

First: In terms of models, we therefore understand *explaining* as the ability to elucidate which factors actually caused a model result. By contrast, the capacity to *understand* the model requires more: This ability consists of being able to anticipate model behavior under counterfactual assumptions.

Figure 2: Explaining and understanding of opaque models.

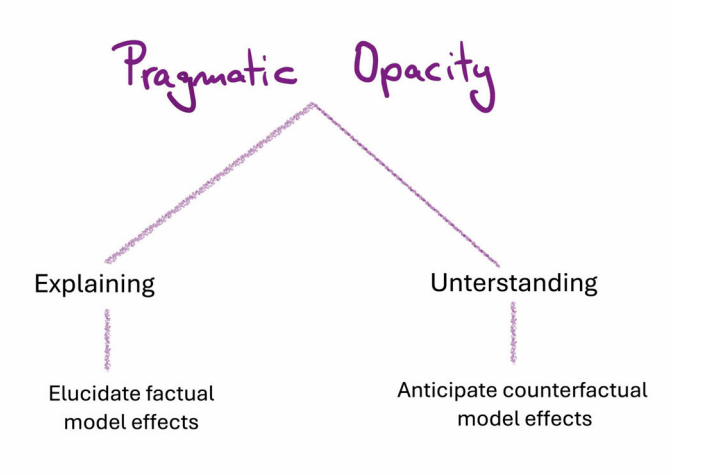


In the case of a credit scoring model, this means that decisions are explained when we are able to identify the decisive factors that led to an actual approval or rejection. We understand the model to the extent that we are able to anticipate, counterfactually, what outcomes the model would produce under altered conditions (e.g., marital status: two instead of three children; place of residence: Bonn instead of Gelsenkirchen; age: 35 instead of 50 years).

Second: With respect to our practices, the aim is to clarify how the use of systems transforms those practices. The capacity to *explain* this consists in articulating how a real system alters our actions; for instance, by reducing or expanding our options, transforming our value orientations, or introducing effects that remain unnoticed

by us.[2] *Understanding* the use of such systems, analogously, refers to the capacity to anticipate, counterfactually, what changed significance a modified system would have for our practices.[3] This latter capacity, in particular, is crucial for the evaluation and design of such systems.

Figure 3: Explaining and understanding how systems affect our practice.



Consider again the case of a credit scoring system: Explanations in this context might consist in identifying why, since the system's introduction, individuals have begun to optimize their creditworthiness in different ways; or why fewer older individuals apply for credit, and how this shift affects economic conditions. The ability to practically understand the use of such systems would mean being able to anticipate, for instance, what consequences a different weighting of certain factors would have for small businesses, families of a given type, and the like. Or what risks might arise for precariously employed individuals if credit were granted to them more easily, and so on.

What remains open, however, is *for whom* models or systems are supposed to be explainable and understandable: individuals, groups, organizations? Or more pre-

2 To our knowledge, there are no papers that directly address pragmatic opacity; nonetheless, there are several lines of thought that point toward aspects that are relevant for this concept. These include works that discuss the use of digital systems operating below the threshold of perception (Nordmann 2008), or that examine the effects of anonymous forms of collectivization ("anonyme Vergemeinschaftung"; Wiegerling et al. 2008: 82).

3 We adopt Beisbart's suggestion here to tie understanding to the role of counterfactuals, even though Beisbart introduced this term with model opacity in mind (Beisbart 2021: 11657).

cisely, with regard to levels of expertise: laypersons, experts – and if so, which ones? More on this later.

4. When are Systems Trustworthy?

Technical systems are trustworthy when they fulfill the relevant values within a given context. At first glance, this statement may seem vague or empty, as it is formulated in abstract terms. However, it is in fact essential to relate trustworthiness to other values that are relevant to consider in specific situations.

Let us walk through this step by step:

Trustworthiness is related to (other) values. This becomes clear when we look at interpersonal relationships. When we trust a person, we expect them to be honest when it matters; to be reliable, kind, courageous, or just when the situation calls for it.

As this shows, trustworthiness is not simply one value among others. Rather, trustworthiness is directed toward the fulfillment of other values. Thus, it is a (higher-order) value of systems that refers to other (first-order) values. Which values these are depends on the situation at hand. Let us compare this to interpersonal trust. If I trust a friend to speak up in the face of an unjust and intimidating colleague, I am relying on her courage. If I trust a friend to pick me up on time so I can catch my flight, it is his reliability that matters in that moment. In such situations, the friend is trustworthy when he acts reliably or, as in the earlier case, courageously.

The same applies *mutatis mutandis* to technical systems. Here, too, trust is directed at trustworthiness, and trustworthiness refers to the relevant values and meta-values (such as reliability, fairness, or privacy, autonomy) and system functions (such as promoting health or enabling prosperity).[4] As a consequence, in order to assess a system's trustworthiness, individuals must evaluate whether the system fulfills the respective first-order values.

Three points, however, must be kept in mind:

(1) The way we talk about values might be understood in a reifying manner – then it may seem as if such language commits us to a reality of values. However, we do not intend to touch on that question here. When we speak of values in what follows, we

4 Hartmann initially describes trust in other people as something that emerges in a relationship that it is not only about trust itself. Trust is part of a practice in which it contributes to realizing values other than the value represented by trust itself for Hartmann (2011: 15–18). The same assumption is often applied to technology. In the following, however, we take a slightly different approach than Hartmann, who develops the relationship between values based on the distinction between instrumental and intrinsic ones.

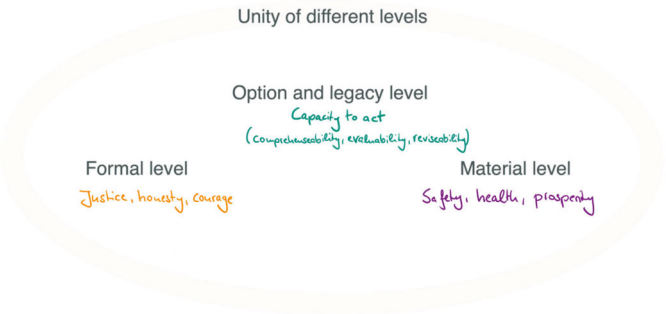
are referring to the practice of evaluating. In this practice of evaluating, we adopt different standpoints. We assess a system in terms of how reliably it produces a certain outcome, how securely it protects data, or how fairly it distributes resources.

(2) The way we spoke about values may suggest that these values operate on the same level. This is by no means the case. We propose distinguishing at least four levels of standpoints, they refer to the various standpoints we can take when assessing systems.

The first level of standpoints involves evaluating systems in *material* terms based on the role or function a system is intended to fulfill, such as prosperity, health, safety, or connectedness. The second level refers to the *formal* assessment of common moral virtues such as justice, courage, or sincerity. This formal perspective often focuses on the material aspect, such as the *just* distribution of *health contributing* resources. What becomes apparent, therefore, is that there is an order among the different levels. This continues with the next aspect as well, for the third level refers to *value enabling values* among those are the option and legacy values. They ensure our capacity to act because they refer to the possibility to revise the course of action and to relate to the effects of a system (which presupposes its explainability and understandability). Thus, the third aspect is connected to the first two. The formal evaluation of systems focuses on the material goods involved such as the fair distribution of resources (e.g., credit) or the sincere and reliable handling of data. The third perspective ensures that systems can be shaped in accordance with the first two viewpoints. This order is finalized on the fourth level: *Trustworthiness* and *responsibility* constitute the *unity* of these values (cf. Kaminski 2021: 396–398).[5] They encompass the other aspects in as much as models can be considered trustworthy and their deployment accountable if and only if they promote, or at least do not violate, material, formal, and value enabling values.

5 Trustworthiness can be understood as a unity, since it integrates the values that are relevant in a given context. To call a person or a system trustworthy is to say that this trustworthiness is constituted by other values that matter in the situation at hand. A person or system is trustworthy, for example, when they are reliable, transparent, or fair in contexts where such qualities are decisive. In this sense, trustworthiness is not an isolated trait but a unifying quality that brings together other values.

Figure 4: The order of viewpoints for the assessment of systems.



The discourse on values may suggest that it is always already clear what these values are and how we determine whether they are being violated. Instead, we adopt the distinction introduced above between models (and the corresponding model opacity) and systems as they are embedded in contexts of action (pragmatic opacity). With respect to models, it indeed seems straightforward to evaluate them by measuring them against certain values such as reliability, safety, or justice. The same, however, does not hold for the deployment of systems in practice. In these cases, it is often not clear which values are relevant; nor is there necessarily a consensus on how to interpret the values in question. Moreover, the very understanding of what counts as a value and which values we consider relevant may itself be altered through the use of these systems. In such situations, what matters is our capacity to understand how systems may transform our practices, so that we may orient ourselves accordingly.

Thesis 3: Trustworthiness refers to a different level of assessing a system. It unifies the lower levels and their viewpoints. If we come to the conclusion that a system is trustworthy, this means that we have evaluated it positively on the other levels – for example, it fulfills its material function of promoting health, and does so in an exemplary formal manner, such as being fair and reliable. In addition, we have access to this system and are able to shape it (option and legacy value).

When a person, community, or organization seeks to assess the trustworthiness of, say, a medical model, they must first evaluate whether and to what extent the model fulfills other values that are relevant within the respective domain, such as health, justice, or reliability. The degree to which these first-order values are promoted or realized by the model determines its trustworthiness. However, if we want to understand how systems transform our practices, it is no longer sufficient to measure models in this way. Two ex-

amples may illustrate this point. A system used in the care of elderly or ill individuals may operate reliably and promote health, yet by primarily treating these individuals as subjects to be relieved of burden, it may diminish their capacity to act (Wiegerling et al. 2008: 75). In doing so, it would violate a basic value (whether such a violation is acceptable or ought to be compensated is a further question). Or consider a system designed to support medical professionals in diagnosis: it may alter their competencies – undermining or developing certain skills, calling into question justified or unjustified self-confidence, and so on. In both cases, even from the perspective of the model, it may not be clear what the relevant values in our practices are, how they should be conceptualized and related to one another, or how they might shift through the use of the system. What we require, then, is an understanding that takes into account the pragmatic opacity of the contexts of action.

5. Responsibility and Digital Governance

'Responsibility' is a normatively underdetermined concept and therefore must always be specified through additional criteria. Responsibility can be conceptualized as follows:

'A is responsible for X to B by appeal to normative standpoint Z.'

There is thus an acting person or institution A who takes responsibility for or is held accountable for an action X. Furthermore, there is a person or institution B to whom the responsibility is owed or by whom it is demanded. This relationship is based on a *normative foundation Z* (e.g., a social norm, a law, etc.) that gives further definition to the responsibility. This basic schema can describe both moral and legal responsibility.[6] The concept of responsibility thus allows for the description of a responsibility relation, but the normative basis for attributing responsibility must always be specified; it does not follow from the concept of 'responsibility' alone. Therefore, appeals to 'responsibility' are always at risk of being used in a purely rhetorical way.

There have been various attempts to define the normative basis of responsibility by tying the attribution of responsibility to the realization of certain values. The approach proposed here, however, assumes that a prior question must be addressed: namely, whether agents are capable of understanding themselves as subjects of responsibility in the first place. It is thus suggested that the attribution of responsibility be expanded to include a reflexive dimension.

6 In one case, it is a matter of moral norms, in the other of legal norms. On this, see Werner 2002: 521et seq.

Thesis 4: For the capacity-based approach, accountability does not (only) mean that systems are measured against certain values. Rather the key question is whether it is possible to relate systems to our practices, norms, values, and goods in an evaluative manner. This, in turn, presupposes digital governance that enables and secures precisely this possibility.

The primary issue is not whether digital systems realize particular values, but whether the acting person A and the entity B – toward whom responsibility is assumed or by whom it is demanded – are capable of determining the normative basis Z. This higher-level dimension of responsibility is necessary because, in the context of digital systems, the normative foundation to which one can appeal is not self-evident. We do not yet know which digital systems are justifiably accountable, and we must first be enabled to make such determinations.[7] This capacity for judgment must be safeguarded and actively shaped and it is precisely at this point that *digital governance* comes into play.

Due to this consideration it becomes evident that an internal connection between responsibility and trust becomes evident. On the one hand, the willingness to engage in a shared understanding of the criteria for attributing responsibility presupposes a basic level of trust. On the other hand, trust can only develop if those criteria are not imposed in an authoritarian manner but remain open to reflexive examination and justification. This connection is of central importance for the functioning of digital systems within democratic and constitutional political orders.

Another key question for digital governance concerns the level at which these capacities (for judgment, reflection, and responsibility) should be developed. Should it be at the level of individuals? If so, should these be affected laypersons? Or should it be organizations such as research institutions or newly established bodies? Given the level of expertise required to understand models and systems, the latter may appear to be the only viable option. However, considering the sheer number and rapid pace at which new systems emerge and become integrated into everyday life, it is equally unrealistic to expect a single central institution to manage this task and assume responsibility for the use of all systems. We therefore argue that digital governance should focus on creating channels of communication and spaces for reflection that connect individuals and organizations.

7 This difference between responsibility and accountability is to be understood in parallel to the difference between acceptance and acceptability. Acceptable does not mean that we accept systems, but that we can decide whether we accept them; this in turn presupposes that we can take a judgmental approach to them. To do this, however, we need to understand how they relate to values and our practice. See Hubig 2007: 115 et seq. et passim.

6. Conclusion

The capacity-approach presented here shifts the perspective on how AI systems are evaluated. Three points stand out:

(1) The central question is no longer (primarily) whether a system fulfills a set of values. Instead, the focus turns to how systems can be evaluated, what capacities are required for such evaluation, and how the development of these capacities can be structured and supported by a system. This is, because these capacities posses high ethical value since they function as preconditions in all value judgements.

(2) Rather than relying mainly on what we have called formal values, evaluation, as a practice, takes place on different levels and through the adoption of various perspectives. In this context, option and legacy perspectives play a particularly prominent role, even though they are largely absent from most approaches in the ethics of technology. Yet they are critically important, since without them (keyword: values enabling values) neither evaluation nor the design of technical systems is possible. Trustworthiness and responsibility are conceived as a unity within the framework of these different levels and perspectives.

3. It becomes evident that the evaluation and design of technical systems is far more demanding than simply checking whether certain (formal) values – such as justice or privacy – are fulfilled. Individuals play a role in this process, but always in conjunction with institutions and organizations that turn the learning processes involved in evaluating and shaping technical systems into a systematic task. AI ethics, therefore, requires AI policy to organize and sustain this effort.

References

Annas, Julia (1995): "Prudence and Morality in Ancient and Modern Ethics", in: Ethics 105(2), pp. 241–257.

Beisbart, Claus (2021): "Opacity Thought Through. On the Intransparency of Computer Simulations", in: Synthese 199(3), pp. 11643–11666.

Benner, Erica (2010): Machiavelli's Ethics, Princeton: Princeton University Press.

Bisconti, Piercosma, et al. (2024): "A Formal Account of AI Trustworthiness. Connecting Intrinsic and Perceived Trustworthiness", in: AIES '24: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society 7(1).

Burrell, Jenna (2016): "How the Machine 'Thinks'. Understanding Opacity in Machine Learning Algorithms", in: Big Data & Society 3(1), pp. 1–12.

Cimakasky, Joseph and Polansky, Ronald (2012): "Descartes' Provisional Morality", in: Pacific Philosophical Quarterly 93(3), pp. 353–372.

Düwell, Marcus (1999): Ästhetische Erfahrung und Moral. Zur Bedeutung des Ästhetischen für die Handlungsspielräume des Menschen, Freiburg: Alber.

Düwell, Marcus, Graf Keyserlingk, Johannes and Richter, Philipp (2025): "Rights-Based Ethics – Outline of an Approach", in: Düwell, Marcus, Graf Keyserlingk, Johannes and Richter. Philipp (eds.), Rights-based Ethics, London/Abingdon: Routledge, pp. 3–32.

Fetic, L. et al. (2020): From Principles to Practice. An Interdisciplinary Framework to Operationalise Ai Ethics, available online: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf.

Floridi, Luciano (2019): "Establishing the Rules for Building Trustworthy AI", in: Nature Machine Intelligence 1(6), pp. 261–262.

Fremstedal, Roe (2018): "Morality and Prudence. A Case for Substantial Overlap and Limited Conflict", in: The Journal of Value Inquiry 52(1), pp. 1–16.

Gewirth, Alan (1978): Reason and Morality, Chicago: University Press.

Hartmann, Martin (2011): Die Praxis des Vertrauens, Berlin: Suhrkamp.

Hubig, Christoph (1990): "Analogie und Ähnlichkeit. Probleme einer theoretischen Begründung vergleichenden Denkens", in: Gerd Jüttemann (ed.), Komparative Kasuistik, Heidelberg: Asanger, pp. 133–142.

Hubig, Christoph (1995): Technik- und Wissenschaftsethik. Ein Leitfaden, Berlin/Heidelberg: Springer.

Hubig, Christoph (2007): Die Kunst des Möglichen II. Ethik der Technik als provisorische Moral, Bielefeld: transcript.

Hubig, Christoph and Richter, Philipp (2015): "Technikethik als Ethik der Ermöglichung des Anwendungsbezuges", in: Ammicht Quinn, Regina and Thomas Potthast (eds.), Ethik in den Wissenschaften, Tübingen: IZEW, pp. 209–214.

Humphreys, Paul (2009): "The Philosophical Novelty of Computer Simulation Methods", in: Synthese 169(3), pp. 615–626.

Kaminski, Andreas, Resch, Michael, and Küster, Uwe (2018): "Mathematische Opazität. Reproduzierbarkeit in der Computersimulation", in: Jahrbuch Technikphilosophie 4, pp. 253–277.

Kant, Immanuel. (2012): Groundwork of the Metaphysics of Morals. Cambridge: University Press.

Kaspar, David (2011): "Can Morality Do Without Prudence?", in: Philosophia 39(2), pp. 311–326.

Longuenesse, Béatrice (2001): Kant and the Capacity to Judge. Sensibility and Discursivity in the Transcendental Analytic of the 'Critique of Pure Reason', Princeton, NJ: Princeton University Press.

Luckner, Andreas (2005): Klugheit, Berlin/New York: DeGruyter.

Luckner, Andreas (2011): "Klugheitsethik", in: Düwell, Marcus et al. (eds.). Handbuch Ethik, 3rd ed., Stuttgart/Weimar: Metzler, pp. 206–217.

Nordmann, Alfred (2008): "Technology Naturalized. A Challenge to Design for the Human Scale", in: Kroes, Peter et al. (eds.), Philosophy and Design. From Engineering to Architecture, Berlin: Springer, pp. 173–184.

Pekka, A.-P. et al. (2018): The European Commission's High-level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy Ai. Working Document for Stakeholders' Consultation. Brussels, pp. 1–37.

Raz, Joseph (1988): The Morality of Freedom, Oxford: University Press.

Richter. Philipp (2017) "Von der 'Wegräumung eines Hindernisses' – Klugheitsethische Topoi als Umsetzungsargumente in den Ethiken des Kantischen Typs", in: Kertscher, Jens and Müller, Jan (eds.), Praxis und zweite Natur. Begründungsfiguren normativer Wirklichkeit in der Diskussion, Münster: Mentis, pp. 187–203.

Richter, Philipp (2018): "Die Unhintergehbarkeit der Reflexion in der anwendungsbezogenen Ethik – eine Positionsbestimmung in klugheitsethisch-topischer Perspektive", in: Müller, Uta, Richter, Philipp and Potthast, Thomas (eds.), Abwägen und Anwenden. Zum 'guten' Umgang mit ethischen Normen und Werten, Tübingen: Narr Francke Attempto, pp. 27–54.

Thiebes, Scott, Lins, Sebastian, and Sunyaev, Ali (2021): "Trustworthy Artificial Intelligence", in: Electronic Markets 31(2), pp. 447–464.

Werner, Micha H. (2002): „Verantwortung", in: Düwell, Marcus, Hübenthal, Christoph, and Werner, Micha H. (eds.), Handbuch Ethik. Stuttgart: Metzler, pp. 521–527.

Wiegerling, Klaus et al. (2008): "Ubiquitärer Computer – Singulärer Mensch", in: Klumpp, Dieter et al. (eds.), Informationelles Vertrauen für die Informationsgesellschaft, Berlin: Springer, pp. 71–84.

Wiggins, David (1975): "Deliberation and Practical Reason", in: Proceedings of the Aristotelian Society 76, pp. 29–51.

Legal Resources

AI Act. 2024: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).